4/prts

TITLE

HOMOLOGS OF MAR-BINDING FILAMENT-LIKE PROTEIN 1 (MFP1)

This application claims the benefit of U.S. Provisional Application No. 60/128,900, filed April 12, 1999.

FIELD OF THE INVENTION

This invention is in the field of plant molecular biology. This invention pertains to nucleic acid fragments encoding proteins that are homolgs to the MAR-binding filament-like protein 1 (MFP1) from tomato. More specifically, this invention pertains to two tobacco MFP1 genes and MFP1 homologs from corn, soybean and rice.

BACKGROUND OF THE INVENTION

The nuclear matrix hypothesis proposes a structural framework for the eukaryotic nucleus that is similar to the cytoskeleton. To date, its best characterized component is the lamina, a filamentous protein network that lines the inner membrane of the nuclear envelope. Major components of the lamina include a group of intermediate-filament (IF) proteins, collectively known as nuclear lamins, that are classified as type A, B, and C (McKeon *et al.*, *Nature* 319:463-468 (1986)). Lamin B is attached to the inner nuclear membrane via a C-terminal C15 farnesyl group (Schafer et al., *Annu. Rev. Genet.* 30:209-237 (1992)), whereas lamins A and C bind to lamin B. Other integral membrane proteins interact with lamin B and most likely stabilize the membrane attachment of lamins (Furukawa et al., *EMBO J.* 14:1626-1636 (1995)). Recent studies have also demonstrated the ability of lamins A and B to bind DNA, suggesting a role for mammalian lamins in anchoring chromatin to the nuclear envelope. The interaction between nuclear envelope, lamina, and chromatin is considered to be of fundamental importance for higher order chromosome organization, as well as the assembly and disassembly of the nuclear envelope during mitosis (Furukawa et al., *EMBO J.* 14:1626-1636 (1995)).

The nuclear matrix is a second structural skeleton that has been biochemically defined as the insoluble component that remains after treatment of isolated nuclei with DNase I and extraction of proteins with high-salt solutions (Berezney et al., *Biochem. Biophys. Res. Comm.* 60:1410-1417 (1974)) or the chaotropic agent lithium diiodosalicylate (Mirkowitch et al., *Cell* 39:223-232 (1984)). Chromatin binds to the nuclear matrix via matrix attachment regions (MARs) in the DNA. MARs are generally AT-rich DNA sequences that are several hundred base pairs long and localized to noncoding regions of the DNA, but often flanking genes (Gasser et al., *Trends Genet.* 3:16-22 (1987)). However,

there is no consensus sequence known for MARs. The significance of structural characteristics for MARs such as DNA bending and a narrow minor groove due to oligo(dA) tracts has been previously proposed. MARs have been shown to increase transcriptional activity of a linked gene and to confer position-independent, copy-number dependent expression in stably transfected cells (Phi-Wan et al., *EMBO J.* 7:655-664 (1988)).

A small number of MAR binding proteins have been identified from animal nuclei, and they are considered to be components of the nuclear matrix (von Kries et al., *Cell* 64:123-135 (1991); Dickinson et al., *Cell* 70:631-645 (1992); Romig et al., *EMBO J.* 11:3431-3440 (1992); Tsutsui et al., *J. Biol. Chem.* 268:12886-12894 (1993); Renz et al., *Nucleic Acids Res.* 24:843-849 (1996); U.S. 5,652,340). In addition, it has been shown that lamins specifically bind to MARs (Luderus et al., *Mol. Cell. Biol.* 14:6297-6305 (1994)). The specific interaction between DNA and the nuclear matrix/nuclear lamina is most likely an important mechanism for long-range gene regulation and higher order chromatin organization (Gasser et al., *Trends Genet.* 3:16-22 (1987)).

Most investigations into structural components of the nucleus have focused on proteins in vertebrates and *Drosophila*, but even in these organisms, our knowledge about the molecular constituents of the nuclear matrix is sparse. Significantly less information is available for other eukaryotes, and in particular for plants. Proteins that are immunologically related to animal IF proteins and lamins have been detected in pea and carrot nuclei (Beven et al., *J. Cell Sci,* (1991) 98 (3), 293-30; McNulty et al., *J. Cell Sci.* 103:407-414 (1992)). Plant nuclear matrix preparations that bind to animal MARs have been reported, suggesting that proteins with similar DNA binding specificities exist in plants as well (Hall et al., *Proc. Natl. Acad. Sci. USA* 88:9320-9324 (1991)).

Effects of MARs on gene expression in plants have been reported, but have been quite variable. In some experimental systems, no reduction of variability but an increase in expression level has been reported (Breyne et al., *Plant Cell* 4:463-471 (1992); Allen et al., *Plant Cell* 5:603-613 (1993); Allen et al., *Plant Cell* 8:899-913 (1996); U.S. 5,773,689). Other authors have found no significant increase in expression level, but a reduction of variability (van der Geest et al., *Plant J.* 6:413-423 (1994); Mlynarova et al., *Plant Cell* 6:417-426 (1994)). It is not clear what causes these observed differences, but they will most probably be due to the fact that MARs establish different molecular interactions, which might either depend on the features of the MAR itself or on the specific molecular environment of the transformed cell/tissue. The routine use of MARs for strategies to improve transgene expression will greatly depend on the

characterization of the proteins involved in DNA-nuclear matrix attachment and the factors responsible for the observed increase in gene expression.

Currently, no sequence information is available for plant lamin-like proteins. However, the cloning of the cDNA for a plant MAR-binding protein, MFP1, from tomato has been reported (Meier et al., *Plant Cell* 8:2105-2115 (1996)). MFP1 has structural features of a filament-like protein and it preferentially binds to MAR DNA sequences from both plants and animals. In contrast to other known MAR binding proteins, MFP1 contains a hydrophobic N-terminal amino acid sequence that might function as a membrane-spanning domain. MFP1, therefore, has features of a novel anchor protein that most likely connects chromatin via MAR DNA with the nuclear envelope and nuclear filament proteins.

In order to routinely use the attachment of transgenes to the nuclear matrix improve gene expression, it will be necessary to further characterize the elements involved in this process and to better understand the underlying mechanisms. Thus, a need exists to identify and characterize additional nuclear matrix proteins. The present invention presents MFP1-like proteins from other plant species. Furthermore, the present invention shows that a single, immunologically related protein of comparable size is present in a variety of higher-plant species, including important crop plants. This invention pertains to the isolation of cDNAs corresponding to two tobacco MFP1 genes and the characterization of the MFP1 gene family in tobacco. The invention also pertains to the identification and partial characterization of EST sequences from corn, soybean and rice encoding MFP1 proteins from these crop species.

SUMMARY OF THE INVENTION

The present invention provides an isolated nucleic acid fragment encoding a plant MFP1 protein selected from the group consisting of: (a) an isolated nucleic acid fragment encoding all or a substantial portion of the amino acid sequence selected from the group consisting of SEQ ID NO:2, SEQ ID NO:4, SEQ ID NO:20, SEQ ID NO:22 and SEQ ID NO:24; (b) an isolated nucleic acid fragment that is substantially similar to an isolated nucleic acid fragment encoding all or a substantial portion of the amino acid sequence selected from the group consisting of SEQ ID NO:2, SEQ ID NO:4, SEQ ID NO:20, SEQ ID NO:22 and SEQ ID NO:24; (c) an isolated nucleic acid molecule that hybridizes with a nucleic acid sequence of (a) or (b) under the following hybridization conditions: 5 x Denhards, 5 x SSPE, 5% SDS, 20 µg/mL salmon sperm DNA at 55 °C; (d) an isolated nucleic acid molecule that hybridizes with a nucleic acid sequence selected from the group consisting of SEQ ID NO:1, SEQ ID NO:3,

SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:19, SEQ ID NO:21 and SEQ ID NO:23 under the following hybridization conditions: 5 x Denhards, 5 x SSPE, 5% SDS, 20 µg/mL salmon sperm DNA at 55 °C; and (e) an isolated nucleic acid fragment that is complementary to (a), (b), (c) or (d).

Additionally the invention provides a nucleic acid fragment, isolated from corn, encoding an MFP1 polypeptide, the polypeptide having at least 40% identity to SEQ ID NO:17, over a length of about 672 amino acids as compared by the Jotun-Hein algorithm.

Similarly the invention provides a nucleic acid fragment, isolated from soybean, encoding an MFP1 polypeptide, the polypeptide having at least 46% identity to SEQ ID NO:17 over a length of 388 amino acids as compared by the Jotun-Hein algorithm.

In another embodiment the invention provide a nucleic acid fragment, isolated, from rice, encoding an MFP1 polypeptide, the polypeptide having at least 39% identity to SEQ ID NO:17 over a length of 107 amino acids as compared by the Jotun-Hein algorithm.

In an alternate embodiment the invention provides an isolated nucleic acid fragment encoding a plant MFP1 polypeptide, the peptide having at least 77% identity to SEQ ID NO:17.

The invention further provides polypeptides encoded by the isolated nucleic acid fragments of the present invention.

In another embodiment the invention provides a chimeric gene comprising the isolated nucleic acid fragment of the present invention operably linked to suitable regulatory sequences.

The invention additionally provides a method of altering the level of expression of a plant MFP1 protein in a host cell comprising: (a) transforming a host cell with the chimeric gene of the present invention and; (b) growing the transformed host cell produced in step (a) under conditions that are suitable for expression of the chimeric gene resulting in production of altered levels of a plant MFP1 protein in the transformed host cell relative to expression levels of an untransformed host cell.

The invention additionally provides transformed host cells comprising the chimeric genes of the present invention.

In an alternate embodiment the invention provides methods of obtaining a nucleic acid fragment encoding all or a substantial portion of the amino acid sequence encoding a plant MFP1 protein using portions of the present nucleic acid sequences as hybridization probes or as primers.

BRIEF DESCRIPTION OF THE DRAWINGS, AND SEQUENCE DESCRIPTIONS

Figure 1 shows a schematic representation of the subfragments E-196 and H-207 that were expressed in *Escherichia coli*.

Figure 2A is a gel showing the immunological identification of MFP1-like proteins in different plant species using the aR50 antibody raised against a Le MFP1 polypeptide.

Figure 2B is a gel showing the immunological identification of MFP1-like proteins in different plant species using the a288 antibody raised against a Le MFP1 polypeptide.

Figure 3 shows the schematic structure of the partial cDNAs isolated from a *Nicotiana tabacum* lambda ZAP cDNA library.

Figure 4A shows the percent identical amino acids in pairwise comparisons of the four MFP1 proteins.

Figure 4B shows the hydrophilicity and secondary structure analysis of LeMFP1, NtMFP1-1 and AtMFP1.

Figure 5 shows the genomic organization of tobacco MFP1.

The invention can be more fully understood from the following detailed description and the accompanying sequence descriptions which form part of this application.

The following sequence descriptions and sequence listings attached hereto comply with the rules governing nucleotide and/or amino acid sequence disclosures in patent applications as set forth in 37 C.F.R. §1.821-1.825. The Sequence Descriptions contain the one letter code for nucleotide sequence characters and the three letter codes for amino acids as defined in conformity with the IUPAC-IYUB standards described in *Nucleic Acids Research* 13:3021-3030 (1985) and in the *Biochemical Journal* 219(2):345-373 (1984) which are herein incorporated by reference. The symbols and format used for nucleotide and amino acid sequence data comply with the rules set forth in 37 C.F.R. §1.822.

SEQ ID NO:1 is the nucleotide sequence for NtMFP1-1.

SEQ ID NO:2 is the deduced amino acid sequence for NtMFP1-1, encoded by SEQ ID NO:1.

SEQ ID NO:3 is the nucleotide sequence for NtMFP1-2.

SEQ ID NO:4 is the deduced amino acid sequence for NtMFP1-2, encoded by SEQ ID NO:3.

SEQ ID NO:5 is the nucleotide sequence which codes for E-196 polypeptide fragment isolated from tomato.

SEQ ID NO:6 is the deduced amino acid sequence for E-196 polypeptide fragment isolated from tomato, encoded by SEQ ID NO:5.

SEQ ID NO:7 is the nucleotide sequence which codes for H-207 polypeptide fragment isolated from tomato.

SEQ ID NO:8 is the amino acid sequence for H-207 polypeptide fragment isolated from tomato, encoded by SEQ ID NO:8.

SEQ ID NO:9 is the nucleotide sequence for the p7-2 fragment isolated from tomato.

SEQ ID NO:10 is the nucleotide sequence for the p1-3 fragment isolated from tomato.

SEQ ID NO:11 is the nucleotide sequence for the T6 fragment isolated from tobacco.

SEQ ID NO:12 is the nucleotide sequence for the T1 fragment isolated from tobacco.

SEQ ID NO:13 is the nucleotide sequence for the T2 fragment isolated from tobacco.

SEQ ID NO:14 is the nucleotide sequence for the T3 fragment isolated from tobacco.

SEQ ID NO:15 is the nucleotide sequence for the PCR1 fragment isolated from tobacco.

SEQ ID NO:16 is the nucleotide sequence for LeMFP1.

SEQ ID NO:17 is the deduced amino acid sequence for LeMFP1, encoded by SEQ ID NO:16.

SEQ ID NO:18 is the nucleotide sequence used as a Southern probe.

SEQ ID NO:19 is the nucleotide sequence comprising the cDNA insert in clone src3c.pk004.m1 encoding a soybean MFP1 (GmMFP1).

SEQ ID NO:20 is the deduced amino acid sequence of the nucleotide sequence comprising the cDNA insert in clone src3c.pk004.m1.

SEQ ID NO:21 is the nucleotide sequence comprising the cDNA insert in clone p0118.chsab48r encoding a corn MFP1.

SEQ ID NO:22 is the deduced amino acid sequence of the nucleotide sequence comprising the cDNA insert in clone p0118.chsab48r.

SEQ ID NO:23 is the nucleotide sequence comprising the cDNA insert in clone rca1n.pk022.a11 encoding a rice MFP1.

SEQ ID NO:24 is the deduced amino acid sequence of the nucleotide sequence comprising the cDNA insert in clone rca1n.pk022.a11.

SEQ ID NO:25 is the nucleotide sequence for PCR primer designed from T3 fragment isolated from tobacco.

SEQ ID NO:26 is the nucleotide sequence for PCR primer designed from T1 fragment isolated from tobacco.

DETAILED DESCRIPTION OF THE INVENTION

The present invention reports the isolation and characterization of cDNAs corresponding to two tobacco MFP1 genes and the isolation and identification of MFP1 EST homologs from corn, soybean and rice. No homologs of MFP1 from tobacco have been described previously. The level of expression of the genes described here can be altered in the plant by methods of cosuppression and overexpression. As they are previously undescribed genes involved in a fundamental cellular mechanism, this can lead to novel developmental phenotypes that might be beneficial for crop growth and development. In addition, if the reduction in expression of one of the genes leads to a growth or developmental defect in the plant, this gene can be used as a novel herbicide target. All isolated proteins can be used as tools to study the plant nuclear matrix, of which no components have been isolated at the molecular level. This can lead to the identification of additional proteins, that can be used as described above. Any related EST sequences can be directly used for the above described applications in crop plants. All of these sequences can be directly used to broaden our understanding of the mechanisms of MAR-matrix interactions and the molecular basis for the described effects on gene expression.

The following definitions are provided for the full understanding of terms and abbreviations used in this specification.

"Polymerase chain reaction" is abbreviated PCR.

"Expressed sequence tag" is abbreviated EST.

"Open reading frame" is abbreviated ORF.

"SDS polyacrylamide gel electrophoresis" is abbreviated SDS-PAGE.

"Amino acid" is abbreviated AA.

"Plaque-forming units" is abbreviated pfus.

"α-Helical" is abbreviated AH.

"Coiled-coil" is abbreviated CC and refers to an amphiphillic α-helical protein structure.

"Hydrophilicity plot" is abbreviated HP.

"Matrix attachment region" is abbreviated MAR. MARs are also known as matrix-associated regions or scaffold-associated (or attachment) regions.

The term "MFP" is an abbreviation for MAR-binding filament-like protein. "MFP1" refers to the MAR-binding filament-like protein having similar characteristics to the protein isolated from tomato as described in Meier et al., *Plant Cell* 8:2105-2115 (1996). "LeMFP1" is the abbreviation for the specific

MFP1 protein isolated from tomato, as set forth in SEQ ID NO:17. "NtMFP1-1" and "NtMFP1-2" are the abbreviations for the first and second MFP1 proteins isolated from tobacco, as set forth in SEQ ID NO:2 and 4 respectively. "GmMFP1" is the abbreviation for the MFP1 protein isolated from soybean, as set forth in SEQ ID NO:20. "ZmMFP1" is the abbreviation for the MFP1 protein isolated from corn, as set forth in SEQ ID NO:22. "OsMFP1" is the abbreviation for the MFP1 protein isolated from rice, as set forth in SEQ ID NO:24. "AtMFP1" is the abbreviation for the MFP1 protein isolated from Arabidopsis, released on (http://genomewww.standard.edu/Arabidopsis/).

The terms "isolated nucleic acid fragment" or "isolated nucleic acid molecule" refer to a polymer of RNA or DNA that is single- or double-stranded, optionally containing synthetic, non-natural or altered nucleotide bases. An isolated nucleic acid fragment or an isolated nulceic acid molecule in the form of a polymer of DNA may be comprised of one or more segments of cDNA, genomic DNA, or synthetic DNA.

The terms "host cell" and "host organism" refer to a cell capable of receiving foreign or heterologous genes and expressing those genes to produce an active gene product. Suitable host cells include microorganisms such as bacteria and fungi, as well as plant cells.

The term "fragment" refers to a DNA or amino acid sequence comprising a subsequence of the nucleic acid sequence or protein of the present invention. However, an active fragment of the present invention comprises a sufficient portion of the protein to maintain activity.

The term "substantially similar" refers to nucleic acid fragments wherein changes in one or more nucleotide bases result in substitution of one or more amino acids, but do not affect the functional properties of the protein encoded by the DNA sequence. "Substantially similar" also refers to nucleic acid fragments wherein changes in one or more nucleotide bases do not affect the ability of the nucleic acid fragment to mediate alteration of gene expression by antisense or co-suppression technology. "Substantially similar" also refers to modifications of the nucleic acid fragments of the present invention such as deletion or insertion of one or more nucleotide bases that do not substantially affect the functional properties of the resulting transcript vis-á-vis the ability to mediate alteration of gene expression by antisense or co-suppression technology or alteration of the functional properties of the resulting protein molecule. It is therefore understood that the invention encompasses more than the specific exemplary sequences.

A "substantial portion" refers to an amino acid or nucleotide sequence which comprises enough of the amino acid sequence of a polypeptide or the

nucleotide sequence of a gene to afford putative identification of that polypeptide or gene, either by manual evaluation of the sequence by one skilled in the art, or by computer-automated sequence comparison and identification using algorithms such as BLAST (Basic Local Alignment Search Tool; Altschul et al., *J. Mol. Biol.* 215:403-410 (1993); see also www.ncbi.nlm.nih.gov/BLAST/). In general, a sequence of ten or more contiguous amino acids or thirty or more nucleotides is necessary in order to putatively identify a polypeptide or nucleic acid sequence as homologous to a known protein or gene. Moreover, with respect to nucleotide sequences, gene-specific oligonucleotide probes comprising 20-30 contiguous nucleotides may be used in sequence-dependent methods of gene identification (e.g., Southern hybridization) and isolation (e.g., *in situ* hybridization of bacterial colonies or bacteriophage plaques). In addition, short oligonucleotides (generally 12 bases or longer) may be used as amplification primers in PCR in order to obtain a particular nucleic acid fragment comprising the primers. Accordingly, a "substantial portion" of a nucleotide sequence comprises enough of the sequence to afford specific identification and/or isolation of a nucleic acid fragment comprising the sequence. The present specification teaches partial or complete amino acid and nucleotide sequences encoding one or more particular plant proteins. The skilled artisan, having the benefit of the sequences as reported herein, may now use all or a substantial portion of the disclosed sequences for the purpose known to those skilled in the art. Accordingly, the present invention comprises the complete sequences as reported in the accompanying Sequence Listing, as well as substantial portions of those sequences as defined above.

For example, it is well known in the art that antisense suppression and co-suppression of gene expression may be accomplished using nucleic acid fragments representing less than the entire coding region of a gene, and by nucleic acid fragments that do not share 100% identity with the gene to be suppressed. Moreover, alterations in a gene that result in the production of a chemically equivalent amino acid at a given site, but do not effect the functional properties of the encoded protein, are well known in the art. Thus, a codon for the amino acid alanine, a hydrophobic amino acid, may be substituted by a codon encoding another less hydrophobic residue, such as glycine, or a more hydrophobic residue, such as valine, leucine, or isoleucine. Similarly, changes which result in substitution of one negatively charged residue for another, such as aspartic acid for glutamic acid, or one positively charged residue for another, such as lysine for arginine, can also be expected to produce a functionally equivalent product. Nucleotide changes which result in alteration of the N-terminal and C-terminal portions of the protein molecule would also not be expected to alter the activity of

the protein. Each of the proposed modifications is well within the routine skill in the art, as is determination of retention of biological activity of the encoded products. Moreover, the skilled artisan recognizes that substantially similar sequences encompassed by this invention are also defined by their ability to hybridize, under stringent conditions (0.1 x SSC, 0.1% SDS, 65 °C), with the sequences exemplified herein. Preferred substantially similar nucleic acid fragments of the present invention are those nucleic acid fragments whose DNA sequences are 80% identical to the DNA sequence of the nucleic acid fragments reported herein. More preferred nucleic acid fragments are 90% identical to the DNA sequence of the nucleic acid fragments reported herein. Most preferred are nucleic acid fragments that are 95% identical to the DNA sequence of the nucleic acid fragments reported herein.

The term "sequence analysis software" refers to any computer algorithm or software program that is useful for the analysis of nucleotide or amino acid sequences. "Sequence analysis software" may be commercially available or independently developed. Typical sequence analysis software will include but is not limited to the GCG suite of programs (Wisconsin Package Version 9.0, Genetics Computer Group (GCG), Madison, WI), BLASTP, BLASTN, BLASTX (Altschul et al., *J. Mol. Biol.* 215:403-410 (1990), and DNASTAR (DNASTAR, Inc. 1228 S. Park St. Madison, WI 53715 USA). Within the context of this application it will be understood that where sequence analysis software is used for analysis, that the results of the analysis will be based on the "default values" of the program referenced, unless otherwise specified. As used herein "default vales" will mean any set of values or parameters which originally load with the software when first initialized.

The term "percent identity" is a relationship between two or more polypeptide sequences or two or more polynucleotide sequences, as determined by comparing the sequences. In the art, "identity" also means the degree of sequence relatedness between polypeptide or polynucleotide sequences, as the case may be, as determined by the match between strings of such sequences. "Identity" and "similarity" can be readily calculated by known methods, including but not limited to those described in: Computational Molecular Biology (Lesk, A. M., ed.) Oxford University Press, New York (1988); Biocomputing: Informatics and Genome Projects (Smith, D. W., ed.) Academic Press, New York (1993); Computer Analysis of Sequence Data, Part I (Griffin, A. M., and Griffin, H. G., eds.) Humana Press, New Jersey (1994); Sequence Analysis in Molecular Biology (von Heinje, G., ed.) Academic Press (1987); and Sequence Analysis Primer (Gribskov, M. and Devereux, J., eds.) Stockton Press, New York (1991).

Preferred methods to determine identity are designed to give the largest match between the sequences tested. Methods to determine identity and similarity are codified in publicly available computer programs. Preferred computer program methods to determine identity and similarity between two sequences include, but are not limited to, the GCG Pileup program found in the GCG program package, using the Needleman and Wunsch algorithm with their standard default values of gap creation penalty=12 and gap extension penalty=4 (Devereux *et al.*, *Nucleic Acids Res.* 12:387-395 (1984)), BLASTP, BLASTN, and FASTA (Pearson *et al.*, *Proc. Natl. Acad. Sci. USA* 85:2444-2448 (1988). The BLASTX program is publicly available from NCBI and other sources (BLAST Manual, Altschul *et al.*, Natl. Cent. Biotechnol. Inf., Natl. Library Med. (NCBI NLM) NIH, Bethesda, Md. 20894; Altschul et al., *J. Mol. Biol.* 215:403-410 (1990); Altschul *et al.*, ÏGapped BLAST and PSI-BLAST: a new generation of protein database search programsÓ, *Nucleic Acids Res.* 25:3389-3402 (1997)). The method to determine percent identity preferred in the present invention is by the method of DNASTAR protein alignment protocol using the Jotun-Hein algorithm (Hein et al., *Methods Enzymol.* 183:626-645 (1990)). Default parameters used for the Jotun-Hein method for alignments are: for multiple alignments, gap penalty=11, gap length penalty=3; for pairwise alignments ktuple=2. As an illustration, for a polynucleotide having a nucleotide sequence with at least 95% identity to a reference nucleotide sequence, it is intended that the nucleotide sequence of the polynucleotide is identical to the reference sequence except that the polynucleotide sequence may include up to five point mutations per each 100 nucleotides of the reference nucleotide sequence. In other words, to obtain a polynucleotide having a nucleotide sequence at least 95% identical to a reference nucleotide sequence, up to 5% of the nucleotides in the reference sequence may be deleted or substituted with another nucleotide, or a number of nucleotides up to 5% of the total nucleotides in the reference sequence may be inserted into the reference sequence. These mutations of the reference sequence may occur at the 5' or 3' terminal positions of the reference nucleotide sequence or anywhere between those terminal positions, interspersed either individually among nucleotides in the reference sequence or in one or more contiguous groups within the reference sequence. Analogously, for a polypeptide having an amino acid sequence having at least 95% "identity" to a reference amino acid sequence, it is intended that the amino acid sequence of the polypeptide is identical to the reference sequence except that the polypeptide sequence may include up to five amino acid alterations per each 100 amino acids of the reference amino acid. In other words, to obtain a polypeptide having an amino acid sequence at least 95% identical to a reference amino acid sequence, up

to 5% of the amino acid residues in the reference sequence may be deleted or substituted with another amino acid, or a number of amino acids up to 5% of the total amino acid residues in the reference sequence may be inserted into the reference sequence. These alterations of the reference sequence may occur at the amino or carboxy terminal positions of the reference amino acid sequence or anywhere between those terminal positions, interspersed either individually among residues in the reference sequence or in one or more contiguous groups within the reference sequence.

"Codon degeneracy" refers to divergence in the genetic code permitting variation of the nucleotide sequence without effecting the amino acid sequence of an encoded polypeptide. Accordingly, the present invention relates to any nucleic acid fragment that encodes all or a substantial portion of present MFP1 proteins as set forth in SEQ ID NO:2, SEQ ID NO:4, SEQ ID NO:20, SEQ ID NO:22 and SEQ ID NO:24. The skilled artisan is well aware of the "codon-bias" exhibited by a specific host cell to use nucleotide codons to specify a given amino acid. Therefore, when synthesizing a gene for improved expression in a host cell, it is desirable to design the gene such that its frequency of codon usage approaches the frequency of preferred codon usage of the host cell.

The term "complementary" is used to describe the relationship between nucleotide bases that are hybridizable to one another. Hence with respect to DNA, adenosine is complementary to thymine and cytosine is complementary to guanine.

A nucleic acid molecule is "hybridizable" to another nucleic acid molecule, such as a cDNA, genomic DNA, or RNA, when a single stranded form of the nucleic acid molecule can anneal to the other nucleic acid molecule under the appropriate conditions of temperature and solution ionic strength. Hybridization and washing conditions are well known and exemplified in Sambrook, J., Fritsch, E. F. and Maniatis, T. Molecular Cloning: A Laboratory Manual, Second Edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor (1989), particularly Chapter 11 and Table 11.1 therein (entirely incorporated herein by reference). The conditions of temperature and ionic strength determine the "stringency" of the hybridization. For preliminary screening for homologous nucleic acids, low stringency hybridization conditions, corresponding to a T_m of 55°, can be used, e.g., 5X SSC, 0.1% SDS, 0.25% milk, and no formamide; or 30% formamide, 5X SSC, 0.5% SDS. Moderate stringency hybridization conditions correspond to a higher T_m, e.g., 40% formamide, with 5X or 6X SSC.

Hybridization requires that the two nucleic acids contain complementary sequences, although depending on the stringency of the hybridization, mismatches between bases are possible. The appropriate stringency for hybridizing nucleic acids depends on the length of the nucleic acids and the degree of complementation, variables well known in the art. The greater the degree of similarity or homology between two nucleotide sequences, the greater the value of T_m for hybrids of nucleic acids having those sequences. The relative stability (corresponding to higher T_m) of nucleic acid hybridizations decreases in the following order: RNA:RNA, DNA:RNA, DNA:DNA. For hybrids of greater than 100 nucleotides in length, equations for calculating T_m have been derived (see Sambrook et al., *supra*, 9.50-9.51). For hybridizations with shorter nucleic acids, i.e., oligonucleotides, the position of mismatches becomes more important, and the length of the oligonucleotide determines its specificity (see Sambrook et al., *supra*, 11.7-11.8). In one embodiment the length for a hybridizable nucleic acid is at least about 10 nucleotides. Preferably a minimum length for a hybridizable nucleic acid is at least about 15 nucleotides; more preferably at least about 20 nucleotides; and most preferably the length is at least 30 nucleotides. Furthermore, the skilled artisan will recognize that the temperature and wash solution salt concentration may be adjusted as necessary according to factors such as length of the probe.

"Synthetic genes" can be assembled from oligonucleotide building blocks that are chemically synthesized using procedures known to those skilled in the art. These building blocks are ligated and annealed to form gene segments which are then enzymatically assembled to construct the entire gene. "Chemically synthesized", as related to a sequence of DNA, means that the component nucleotides were assembled *in vitro*. Manual chemical synthesis of DNA may be accomplished using well established procedures, or automated chemical synthesis can be performed using one of a number of commercially available machines. Accordingly, the genes can be tailored for optimal gene expression based on optimization of nucleotide sequence to reflect the codon bias of the host cell. The skilled artisan appreciates the likelihood of successful gene expression if codon usage is biased towards those codons favored by the host. Determining preferred codons can be based on a survey of genes derived from the host cell where sequence information is available.

"Gene" refers to a nucleic acid fragment that expresses a specific protein, including regulatory sequences preceding (5' non-coding sequences) and following (3' non-coding sequences) the coding sequence. "Native gene" refers to a gene as found in nature with its own regulatory sequences. "Chimeric gene"

refers to any gene, not a native gene, comprising regulatory and coding sequences that are not found together in nature. Accordingly, a chimeric gene may comprise regulatory sequences and coding sequences that are derived from different sources, or regulatory sequences and coding sequences derived from the same source, but arranged in a manner different than that found in nature. "Endogenous gene" refers to a native gene in its natural location in the genome of an organism. A "foreign" gene refers to a gene not normally found in the host organism, but which is introduced into the host organism by gene transfer. Foreign genes can comprise native genes inserted into a non-native organism, or chimeric genes. A "transgene" is a gene that has been introduced into the genome by a transformation procedure.

"Coding sequence" refers to a DNA sequence that codes for a specific amino acid sequence. "Regulatory sequences" refer to nucleotide sequences located upstream (5' non-coding sequences), within, or downstream (3' non-coding sequences) of a coding sequence, and which influence the transcription, RNA processing or stability, or translation of the associated coding sequence. Regulatory sequences may include promoters, translation leader sequences, introns, and polyadenylation recognition sequences.

"Promoter" refers to a DNA sequence capable of controlling the expression of a coding sequence or functional RNA. In general, a coding sequence is located 3' to a promoter sequence. The promoter sequence consists of proximal and more distal upstream elements, the latter elements often referred to as enhancers. Accordingly, an "enhancer" is a DNA sequence which can stimulate promoter activity and may be an innate element of the promoter or a heterologous element inserted to enhance the level or tissue-specificity of a promoter. Promoters may be derived in their entirety from a native gene, or be composed of different elements derived from different promoters found in nature, or even comprise synthetic DNA segments. It is understood by those skilled in the art that different promoters may direct the expression of a gene in different tissues or cell types, or at different stages of development, or in response to different environmental conditions. Promoters which cause a gene to be expressed in most cell types at most times are commonly referred to as "constitutive promoters". New promoters of various types useful in plant cells are constantly being discovered; numerous examples may be found in the compilation by Okamuro and Goldberg, (*Biochemistry of Plants* 15:1-82 (1989)). It is further recognized that since in most cases the exact boundaries of regulatory sequences have not been completely defined, DNA fragments of different lengths may have identical promoter activity.

The "translation leader sequence" refers to a DNA sequence located between the promoter sequence of a gene and the coding sequence. The translation leader sequence is present in the fully processed mRNA upstream of the translation start sequence. The translation leader sequence may affect processing of the primary transcript to mRNA, mRNA stability or translation efficiency. Examples of translation leader sequences have been described (Turner et al., *Mol. Biotech.* 3:225 (1995)).

The "3' non-coding sequences" refer to DNA sequences located downstream of a coding sequence and include polyadenylation recognition sequences and other sequences encoding regulatory signals capable of affecting mRNA processing or gene expression. The polyadenylation signal is usually characterized by affecting the addition of polyadenylic acid tracts to the 3' end of the mRNA precursor. The use of different 3' non-coding sequences is exemplified by Ingelbrecht et al. (*Plant Cell* 1:671-680 (1989)).

"RNA transcript" refers to the product resulting from RNA polymerase-catalyzed transcription of a DNA sequence. When the RNA transcript is a perfect complementary copy of the DNA sequence, it is referred to as the primary transcript or it may be a RNA sequence derived from posttranscriptional processing of the primary transcript and is referred to as the mature RNA. "Messenger RNA" (mRNA) refers to the RNA that is without introns and that can be translated into protein by the cell. "cDNA" refers to a double-stranded DNA that is complementary to and derived from mRNA. "Sense" RNA refers to RNA transcript that includes the mRNA and so can be translated into protein by the cell. "Antisense RNA" refers to a RNA transcript that is complementary to all or part of a target primary transcript or mRNA and that blocks the expression of a target gene (U.S. 5,107,065). The complementarity of an antisense RNA may be with any part of the specific gene transcript, i.e., at the 5' non-coding sequence, 3' non-coding sequence, introns, or the coding sequence. "Functional RNA" refers to antisense RNA, ribozyme RNA, or other RNA that is not translated yet has an effect on cellular processes.

The term "operably-linked" refers to the association of nucleic acid sequences on a single nucleic acid fragment so that the function of one is affected by the other. For example, a promoter is operably-linked with a coding sequence when it affects the expression of that coding sequence (i.e., that the coding sequence is under the transcriptional control of the promoter). Coding sequences can be operably-linked to regulatory sequences in sense or antisense orientation.

The term "expression" refers to the transcription and stable accumulation of sense (mRNA) or antisense RNA derived from the nucleic acid fragment of the

invention. Expression may also refer to translation of mRNA into a polypeptide. "Antisense inhibition" refers to the production of antisense RNA transcripts capable of suppressing the expression of the target protein. "Overexpression" refers to the production of a gene product in transgenic organisms that exceeds levels of production in normal or non-transformed organisms. "Co-suppression" refers to the production of sense RNA transcripts capable of suppressing the expression of identical or substantially similar foreign or endogenous genes (U.S. 5,231,020).

"Altered levels" refers to the production of gene product(s) in organisms in amounts or proportions that differ from that of normal or non-transformed organisms.

"Mature" protein refers to a post-translationally processed polypeptide; i.e., one from which any pre- or propeptides present in the primary translation product have been removed. "Precursor" protein refers to the primary product of translation of mRNA; i.e., with pre- and propeptides still present. Pre- and propeptides may be but are not limited to intracellular localization signals.

A "chloroplast transit peptide" is an amino acid sequence which is translated in conjunction with a protein and directs the protein to the chloroplast or other plastid types present in the cell in which the protein is made. "Chloroplast transit sequence" refers to a nucleotide sequence that encodes a chloroplast transit peptide. A "signal peptide" is an amino acid sequence which is translated in conjunction with a protein and directs the protein to the secretory system (Chrispeels, J. J., (1991) *Ann. Rev. Plant Phys. Plant Mol. Biol. 42*:21-53). If the protein is to be directed to a vacuole, a vacuolar targeting signal (*supra*) can further be added, or if to the endoplasmic reticulum, an endoplasmic reticulum retention signal (*supra*) may be added. If the protein is to be directed to the nucleus, any signal peptide present should be removed and instead a nuclear localization signal included (Raikhel (1992) *Plant Phys.100*:1627-1632).

"Transformation" refers to the transfer of a nucleic acid fragment into the genome of a host organism, resulting in genetically stable inheritance. Host organisms containing the transformed nucleic acid fragments are referred to as "transgenic" organisms. Examples of methods of plant transformation include Agrobacterium-mediated transformation (De Blaere et al., *Meth. Enzymol.* 143:277 (1987)) and particle-accelerated or "gene gun" transformation technology (Klein *et al.*, *Nature, London* 327:70-73 (1987); U.S. 4,945,050).

Standard recombinant DNA and molecular cloning techniques used herein are well known in the art and are described more fully in Sambrook, J., Fritsch,

E.F. and Maniatis, T. *Molecular Cloning: A Laboratory Manual*; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, 1989 (hereinafter "Maniatis").

Novel MFP1-binding proteins, have been isolated from tobacco, corn, soybean and rice. Comparison of their random cDNA sequences to the GenBank database using the BLAST and DNASTAR algorithms, well known to those skilled in the art, revealed that these proteins have no significant homologies to other known proteins, other than MFP1 proteins. The nucleotide sequences of the present MFP1 cDNA are provided in SEQ ID NO:1, SEQ ID NO:3, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:19, SEQ ID NO:21 and SEQ ID NO:23. Other MFP1 genes and proteins from other plants can now be identified by comparison of random cDNA sequences to the present MFP1 sequences provided herein.

Comparison of the instant MFP1 base deduced amino acid sequences to the only published sequence of this kind (LeMFP1, Meier *et al.*, *Plant Cell* 8:2105-2115 (1996); SEQ ID NO:17 and 18) show a variation of homology of about 39% identity (rice, SEQ ID NO:24) over a length of 107 amino acids to about 77% identity for tobacco (SEQ ID NO:2 and 4) as compared by the Jotun-Hein alignment algorithm (Hein et al., *Methods Enzymol.* 183:626-645 (1990)).

Accordingly preferred polypeptides of the instant invention are those plant proteins which are at least 77% identical to the amino acid sequence as set forth in SEQ ID 17. More preferred amino acid fragments are at least about 80%-90% identical to the sequences herein. Most preferred are nucleic acid fragments that are at least 95% identical to the amino acid fragments reported herein. Similarly, preferred nucleic acid sequences are those encoding MFP1 binding proteins and which are at least 80% identical to the nucleic acid sequences of reported herein. More preferred nucleic acid fragments are at least 90% identical to the sequences herein. Most preferred are nucleic acid fragments that are at least 95% identical to the nucleic acid fragments reported herein.

Similarly preferred polypeptides are those isolated from corn which are at least 40% identical to the polypeptide of SEQ ID NO:17 over a length of about 672 amino acids as compare by the Jotun-Hein alignment algorithm (Hein et al., *supra*). Other preferred polypeptides are those isolated from rice which are at least 39% identical to the polypeptide of SEQ ID NO:17 over a length of about 107 amino acids as compare by the Jotun-Hein alignment algorithm (Hein et al., *supra*). Additionally preferred polypeptides are those isolated from soybean which are at least 46% identical to the polypeptide of SEQ ID NO:17 over a

length of about 388 amino acids as compare by the Jotun-Hein alignment algorithm (Hein et al., *supra*).

The nucleic acid fragments of the present invention may be used to isolate cDNAs and genes encoding a homologous MFP1 proteins from the same or other plant species. Isolating homologous genes using sequence-dependent protocols is well known in the art. Examples of sequence-dependent protocols include, but are not limited to, methods of nucleic acid hybridization and methods of DNA and RNA amplification as exemplified by various uses of nucleic acid amplification technologies (e.g., polymerase chain reaction (PCR) or ligase chain reaction).

For example, other MFP1 genes, either as cDNAs or genomic DNAs, could be isolated directly by using all or a portion of the present nucleic acid fragments as DNA hybridization probes to screen libraries from any desired plant using methodology well known to those skilled in the art. Specific oligonucleotide probes based upon the present MFP1 sequences can be designed and synthesized by methods known in the art (Maniatis, *supra*). Moreover, the entire sequences can be used directly to synthesize DNA probes by methods known to the skilled artisan such as random primers, DNA labeling, nick translation, or end-labeling techniques, or RNA probes using available *in vitro* transcription systems. In addition, specific primers can be designed and used to amplify a part of or full-length of the present sequences. The resulting amplification products can be labeled directly during amplification reactions or labeled after amplification reactions, and used as probes to isolate full length cDNA or genomic fragments under conditions of appropriate stringency.

In addition, two short segments of the present nucleic acid fragment may be used in PCR protocols to amplify longer nucleic acid fragments encoding homologous MFP1 genes from DNA or RNA. The polymerase chain reaction may also be performed on a library of cloned nucleic acid fragments wherein the sequence of one primer is derived from the present nucleic acid fragments, and the sequence of the other primer takes advantage of the presence of the polyadenylic acid tracts to the 3' end of the mRNA precursor encoding plant MFP1.

Alternatively, the second primer sequence may be based upon sequences derived from the cloning vector. For example, the skilled artisan can follow the RACE protocol (Frohman et al., *Proc. Natl. Acad. Sci., USA* 85:8998 (1988)) to generate cDNAs by using PCR to amplify copies of the region between a single point in the transcript and the 3' or 5' end. Primers oriented in the 3' and 5' directions can be designed from the present sequences. Using commercially available 3' RACE or 5' RACE systems (BRL), specific 3' or 5' cDNA fragments can be isolated (Ohara et al., *Proc. Natl. Acad. Sci., USA* 86:5673 (1989); Loh

et al., *Science* 243:217 (1989)). Products generated by the 3' and 5' RACE procedures can be combined to generate full-length cDNAs (Frohman et al., *Techniques* 1:165 (1989)).

Finally, availability of the present nucleotide and deduced amino acid sequences facilitates immunological screening of cDNA expression libraries. Synthetic peptides representing portions of the present amino acid sequences may be synthesized. These peptides can be used to immunize animals to produce polyclonal or monoclonal antibodies with specificity for peptides or proteins comprising the amino acid sequences. These antibodies can be then be used to screen cDNA expression libraries to isolate full-length cDNA clones of interest (Lerner et al., *Adv. Immunol.* 36:1 (1984); Maniatis, *supra*).

The nucleic acid fragments of the present invention may also be used to create transgenic plants in which the present MFP1 protein is present at higher or lower levels than normal. Alternatively, in some applications, it might be desirable to express the present MFP1 protein in specific plant tissues and/or cell types, or during developmental stages in which they would normally not be encountered.

Overexpression of the present MFP1 may be accomplished by first constructing a chimeric gene in which the MFP1coding region is operably-linked to a promoter capable of directing expression of a gene in the desired tissues at the desired stage of development. For reasons of convenience, the chimeric gene may comprise promoter sequences and translation leader sequences derived from the same genes. 3' Non-coding sequences encoding transcription termination signals must also be provided. The present chimeric genes may also comprise one or more introns in order to facilitate gene expression.

Plasmid vectors comprising the present chimeric genes can then be constructed. The choice of a plasmid vector depends upon the method that will be used to transform host plants. The skilled artisan is well aware of the genetic elements that must be present on the plasmid vector in order to successfully transform, select and propagate host cells containing the chimeric gene. The skilled artisan will also recognize that different independent transformation events will result in different levels and patterns of expression (Jones et al., *EMBO J.* 4:2411-2418 (1985); De Almeida et al., *Mol. Gen. Genetics* 218:78-86 (1989)), and thus that multiple events must be screened in order to obtain lines displaying the desired expression level and pattern. Such screening may be accomplished by Southern analysis of DNA, Northern analysis of mRNA expression, Western analysis of protein expression, or phenotypic analysis.

For some applications it may be useful to direct the MFP1 protein to different cellular compartments or to facilitate their secretion from the cell. The chimeric genes described above may be further modified by the addition of appropriate intracellular or extracellular targeting sequence to their coding regions. These include chloroplast transit peptides (Keegstra et al., *Cell* 56:247-253 (1989), signal sequences that direct proteins to the endoplasmic reticulum (Chrispeels et al., *Ann. Rev. Plant Phys. Plant Mol.* 42:21-53 (1991), and nuclear localization signal (Raikhel et al., *Plant Phys.* 100:1627-1632 (1992). While the references cited give examples of each of these, the list is not exhaustive and more targeting signals of utility may be discovered in the future.

It may also be desirable to reduce or eliminate expression of the MFP1 genes in plants for some applications. In order to accomplish this, chimeric genes designed for antisense or co-suppression of MFP1 can be constructed by linking the genes or gene fragments encoding parts of these enzymes to plant promoter sequences. Thus, chimeric genes designed to express antisense RNA for all or part of MFP1 can be constructed by linking the MFP1 genes or gene fragments in reverse orientation to plant promoter sequences. The co-suppression or antisense chimeric gene constructs could be introduced into plants via well known transformation protocols wherein expression of the corresponding endogenous genes are reduced or eliminated.

The present MFP1 proteins may be produced in heterologous host cells, particularly in the cells of microbial hosts, and can be used to prepare antibodies to the proteins by methods well known to those skilled in the art. The antibodies would be useful for detecting the present MFP1 protein *in situ* in cells or *in vitro* in cell extracts. Preferred heterologous host cells for production of the present MFP1 protein are microbial hosts. Microbial expression systems and expression vectors containing regulatory sequences that direct high level expression of foreign proteins are well known to those skilled in the art. Any of these could be used to construct a chimeric gene for production of the present MFP1. This chimeric gene could then be introduced into appropriate microorganisms via transformation to provide high level expression of the present MFP1 protein.

Microbial host cells suitable for the expression of the present MFP1 proteins include any cell capable of expression of the chimeric genes encoding these proteins. Such cells will include both bacteria and fungi including, for example, the yeasts (e.g., *Aspergillus*, *Saccharomyces*, *Pichia*, *Candida*, and *Hansenula*), members of the genus *Bacillus* as well as the enteric bacteria (e.g., *Escherichia*, *Salmonella*, and *Shigella*). Methods for the transformation of such hosts and the expression of foreign proteins are well known in the art and

examples of suitable protocols may be found in Manual of Methods for General Bacteriology (Gerhardt et al., eds., American Society for Microbiology, Washington, DC. (1994)) or in Biotechnology: A Textbook of Industrial Microbiology, Second Edition, Brock, T. D., Sinauer Associates, Inc., Sunderland, MA (1989)).

Vectors or cassettes useful for transforming suitable microbial host cells are well known in the art. Typically the vector or cassette contains sequences directing transcription and translation of the relevant gene, a selectable marker, and sequences allowing autonomous replication or chromosomal integration. Suitable vectors comprise a region 5' of the gene which harbors transcriptional initiation controls and a region 3' of the DNA fragment which controls transcriptional termination. It is most preferred when both control regions are derived from genes homologous to the transformed host cell, although such control regions need not be derived from the genes native to the specific species chosen as a production host.

Initiation control regions or promoters useful to drive expression of the genes encoding the MFP1 proteins in the desired host cell are numerous and familiar to those skilled in the art. Virtually any promoter capable of driving these genes is suitable for the present invention including but not limited to CYC1, HIS3, GAL1, GAL10, ADH1, PGK, PHO5, GAPDH, ADC1, TRP1, URA3, LEU2, ENO, TPI (useful for expression in *Saccharomyces*); AOX1 (useful for expression in *Pichia*); and lac, trp, lP_L, lP_R, T7, tac, and trc (useful for expression in *E. coli*). Termination control regions may also be derived from various genes native to the preferred hosts. Optionally, a termination site may be unnecessary; however, it is most preferred if included.

Additionally, the present MFP1 proteins can be used as targets to facilitate the design and/or identification of inhibitors of MFP1 that may be useful as herbicides or fungicides. This could be achieved either through the rational design and synthesis of potent functional inhibitors that result from structural and/or mechanistic information that is derived from the purified present plant proteins, or through random *in vitro* screening of chemical libraries. It is anticipated that significant *in vivo* inhibition of any of the MFP1 proteins described herein may severely cripple cellular metabolism and likely result in plant (or fungal) death.

All or a portion of the nucleic acid fragments of the present invention may also be used as probes for genetically and physically mapping the genes that they are a part of, and as markers for traits linked to expression of the present MFP1. Such information may be useful in plant breeding in order to develop lines with

desired phenotypes. For example, the present nucleic acid fragments may be used as restriction fragment length polymorphism (RFLP) markers. Southern blots (Maniatis, *supra*) of restriction-digested plant genomic DNA may be probed with the nucleic acid fragments of the present invention. The resulting banding patterns may then be subjected to genetic analyses using computer programs such as MapMaker (Lander et al., *Genomics* 1:174-181 (1987)) in order to construct a genetic map. In addition, the nucleic acid fragments of the present invention may be used to probe Southern blots containing restriction endonuclease-treated genomic DNAs of a set of individuals representing parent and progeny of a defined genetic cross. Segregation of the DNA polymorphisms is noted and used to calculate the position of the present nucleic acid sequence in the genetic map previously obtained using this population (Botstein et al., *Am. J. Hum. Genet.* 32:314-331 (1980)).

The production and use of plant gene-derived probes for use in genetic mapping is described by Bernatzky et al. (*Plant Mol. Biol. Reporter* 4:37-41 (1986)). Numerous publications describe genetic mapping of specific cDNA clones using the methodology outlined above or variations thereof. For example, F2 intercross populations, backcross populations, randomly mated populations, near isogenic lines, and other sets of individuals may be used for mapping. Such methodologies are well known to those skilled in the art.

Nucleic acid probes derived from the present nucleic acid sequences may also be used for physical mapping (i.e., placement of sequences on physical maps; *see* Hoheisel et al., Nonmammalian Genomic Analysis: A Practical Guide, pp. 319-346, Academic Press (1996), and references cited therein).

In another embodiment, nucleic acid probes derived from the present nucleic acid sequence may be used in direct fluorescence *in situ* hybridization (FISH) mapping. Although current methods of FISH mapping favor use of large clones (several to several hundred kb), improvements in sensitivity may allow performance of FISH mapping using shorter probes.

A variety of nucleic acid amplification-based methods of genetic and physical mapping may be carried out using the present nucleic acid sequences. Examples include allele-specific amplification (Kazazian et al., *J. Lab. Clin. Med.* 114:95-96 (1989)), polymorphism of PCR-amplified fragments (CAPS; Sheffield et al., *Genomics* 16:325-332 (1993)), allele-specific ligation (Landegren et al., *Science* 241:1077-1080 (1988)), nucleotide extension reactions (Sokolov et al., *Nucleic Acid Res.* 18:3671 (1990)), Radiation Hybrid Mapping (Walter et al., *Nature Genetics* 7:22-28 (1997)) and Happy Mapping (Dear *et al.*, *Nucleic Acid Res.* 17:6795-6807 (1989)). For these methods, the sequence of a nucleic acid

fragment is used to design and produce primer pairs for use in the amplification reaction or in primer extension reactions. The design of such primers is well known to those skilled in the art. In methods using PCR-based genetic mapping, it may be necessary to identify DNA sequence differences between the parents of the mapping cross in the region corresponding to the present nucleic acid sequence. This, however, is generally not necessary for mapping methods.

Loss of function-mutant phenotypes may be identified for the present cDNA clones either by targeted gene disruption protocols or by identifying specific mutants for these genes contained in a maize population carrying mutations in all possible genes (Ballinger et al., *Proc. Natl. Acad. Sci. USA* 86:9402 (1989); Koes et al., *Proc. Natl. Acad. Sci. USA* 92:8149 (1995); Bensen et al., *Plant Cell* 7:75 (1995)). The latter approach may be accomplished in two ways. First, short segments of the present nucleic acid fragments may be used in polymerase chain reaction protocols in conjunction with a mutation tag sequence primer on DNAs prepared from a population of plants in which Mutator transposons or some other mutation-causing DNA element has been introduced (see Bensen, *supra*). The amplification of a specific DNA fragment with these primers indicates the insertion of the mutation tag element in or near the plant gene encoding the MFP1 protein. Alternatively, the present nucleic acid fragment may be used as a hybridization probe against PCR amplification products generated from the mutation population using the mutation tag sequence primer in conjunction with an arbitrary genomic site primer, such as that for a restriction enzyme site-anchored synthetic adaptor. With either method, a plant containing a mutation in the endogenous gene encoding a MFP1 protein can be identified and obtained. This mutant plant can then be used to determine or confirm the natural function of the MFP1 gene product.

The present invention is further defined in the following Examples, in which all parts and percentages are by weight and degrees are Celsius, unless otherwise stated. It should be understood that these Examples, while indicating preferred embodiments of the invention, are given by way of illustration only. From the above discussion and these Examples, one skilled in the art can ascertain the essential characteristics of this invention, and without departing from the spirit and scope thereof, can make various changes and modifications of the invention to adapt it to various usage and conditions.

EXAMPLES

GENERAL METHODS

Standard recombinant DNA and molecular cloning techniques used here are well known in the art and are described by Sambrook et al. (1989), J., Fritsch,

E. F. and Maniatis, T. Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1989 (hereinafter "Maniatis"); and by T. J. Silhavy, M. L. Bennan, and L. W. Enquist, Experiments with Gene Fusions, Cold Spring Harbor Laboratory Press, Cold Spring, N.Y. (1984) and by Ausubel et al., Current Protocols in Molecular Biology, pub. by Greene Publishing Assoc. and Wiley-Interscience (1987).

Nucleotide and amino acid percent identity and similarity comparisons were made using the DNASTAR suite of programs, applying default parameters unless indicated otherwise.

The meaning of abbreviations is as follows: "sec" means second(s), "min"means minute(s), "h" means hour(s), "d" means day(s), "μL" means microliter, "mL" means milliliters, "L" means liters, "mM" means millimolar, "M" means molar, "mmol" means millimole(s).

Plant material and growth conditions:

Tobacco, tomato, soybean, rice, corn, wheat and *Arabidopsis thaliana* were grown in soil in a growth chamber with a 12 h 24 °C light cycle followed by a 12 h, 20 °C dark cycle.

EXAMPLE 1

Isolation Of Total Protein

Total protein extracts were prepared from leaf tissues. 100 mg aliquots of tissue were ground to a fine powder with mortar and pestle in liquid nitrogen, resuspended in 0.5 mL extraction buffer (62.5 mM Tris-Cl, pH 6.8, 20% glycerol, 4% SDS, and 1.4 M β-mercaptoethanol) and incubated at 70 °C for 10 min. The debri was removed by centrifugation at 15,000 rpm for 10 min at 4 °C. The supernatants were removed to new tubes, frozen in liquid nitrogen, and stored at -80 °C.

EXAMPLE 2

Protein Expression, Purification And Antibody Production

pRSETC-MFP1-EcoRI (containing the coiled-coil domain) and pRSETA-MFP1-HincII (containing the DNA binding domain), the expression vectors for E-196 (SEQ ID NO:6) coded by SEQ ID NO:5 and H-207 (SEQ ID NO:8) coded by SEQ ID NO:7 fragments, respectively, have been described previously (Meier *et al.*, *Plant Cell* 8:2105-2115 (1996)). Figure 1 shows a representation of the subfragments E-196 and H-207 that were expressed in *Escherichia coli*. Filled bars indicate α-helical regions, open bars indicate hydrophobic domains. The shaded box marks the DNA-binding domain. Numbers indicate the position of the first and last amino acid of each subfragment. Expression of recombinant fusion proteins containing an N-terminal 6-histidine

tag fused to the protein subfragments E-196 (SEQ ID NO:6) and H-207 (SEQ ID NO:8) was induced by isopropyl-D-thiogalactoside in *Escherichia coli* BL21 cells according to the Qiagen protein expression manual (Qiagen, Chatsworth, CA). The amount of fusion protein present in the different total *E. coli* protein extracts was determined by immunoblotting (Maniatis) with a monoclonal antibody directed against the T7 tag (Novagen, Madison, Wisconsin). The expressed proteins were purified by nickel-affinity chromatography (Qiagen, Chatsworth, CA), followed by SDS PAGE. The bands corresponding to the fusion proteins (~1 mg each) were excised from the gel, ground and used to raise two rabbit antisera (a288 against E-196 (SEQ ID NO:6) and aR50 against H-207 (SEQ ID NO:8)). Polyclonal antibodies were produced in rabbits by Eurogentech, Belgium, using the company's standard immunization protocols. The a288 antibody has been described previously (Meier et al., *Plant Cell* 8:2105-2115 (1996)).

EXAMPLE 3

Immuno-detection Of MFP1 Related Proteins In A Variety Of Higher Plant Species

A 1:3000 dilution of a288 or aR50 antiserum, and a 1:5000 dilution of horseradish peroxidase-coupled anti-rabbit secondary antibody (Amersham, Buckinghamshire, England) were used to perform immunoblot analyses (Maniatis). Enhanced chemiluminescence detection was performed using an ECL detection kit as described by the manufacturer (Amersham Buckinghamshire, England).

a288 and aR50 polyclonal antibodies were then used to detect proteins with antigenic similarity to MFP1 in other plant species (Figure 2). Total protein extracts were prepared from mature leaf tissues of tomato (*Lycopersicon esculentum L.*), tobacco (*Nicotiana tabacum L.*), *Arabidopsis thaliana*, soybean (*Glycine max L.*), rice (*Oryza sativa L.*), wheat (*Triticum aestivum L.*), and corn (*Zea mays L.*) as described above. Equal amounts of total protein, as determined by Coomassie Brilliant Blue staining of a replica protein gel, were probed in immunoblot experiments with aR50 (Figure 2A) and a288 (Figure 2B) polyclonal antibodies. The arrow indicates the position of the MFP1-like proteins of approximately equal size in both panels. The position of molecular weight markers is indicated. aR50 antibody detects a single protein of slightly variable size in all species tested. A second band of higher molecular weight (asterick in Figure 2A) was only occasionally observed in tomato and tobacco extracts (tobacco not shown in Figures 2A-B) and might represent an aggregate of MFP1. In contrast, a288 antibody only detected a protein of about 80 kD in tomato and

tobacco extracts, suggesting that the DNA-binding domain of MFP1 is more highly conserved than the part of the coiled-coil domain present on fragment E-196.

Together, these data indicate that a protein of similar size, containing a related DNA-binding domain, is conserved among higher plant species and that the highest degree of similarity to tomato (LeMFP1) among the plants investigated can be expected in tobacco.

EXAMPLE 4

Cloning and Characterization Of Several Tobacco MFP1 cDNAs Corresponding To Two Tobacco MFP1 Proteins

Example 4 describes the cloning and characterization of two, distinct MFP1 proteins from tobacco.

Isolation Of cDNA By Hybridization

The cDNAs encoding tobacco MFP1 were cloned and characterized. An oligo-dT-primed lambda-ZAP cDNA library made from *Nicotiana tabacum var.* SR1 leaf tissue was purchased from Stratagene (La Jolla, CA). The library was screened in a DNA-hybridization screen according to Maniatis with a 1.6 kb partial cDNA clone representing the 3' 2/3 of the tomato homolog of MFP1, LeMPF1 cDNA (p7 2; SEQ ID NO:9) (Meier et al., *Plant Cell* 8:2105-2115 (1996)) or a 1.0 kb 5' partial LeMFP1 cDNA clone (p1-3; SEQ ID NO:10) (Meier et al., *Plant Cell* 8:2105-2115 (1996)). Hybridization conditions were 5 x Denhards (Maniatis), 5 x SSPE (Maniatis), 5% SDS, 20 μg/mL salmon sperm DNA at 55 °C. Washes were performed at high stringency (0.1 x SSC, 0.1 % SDS at 65 °C). Positive plaques were detected by autoradiography and carried through two subsequent rounds of purification, as described above. *In vivo* excision of positive phage was performed according to the manufacturer's protocol (Stratagene, La Jolla, CA).

Sequencing

In the first screen two positive plaque-forming units (pfus) were detected among approximately 600,000 pfus. After *in vivo* excision, sequence analysis of the two excised cDNAs (T6 (SEQ ID NO:11) and T1 (SEQ ID NO:12)) showed that they represented 1103 bp and 912 bp C-terminal MFP1 sequence, respectively (Figure 3). DNA sequencing was carried out using an ABI Model 377 Sequencer (Perkin Elmer-ABI, Foster City, CA). Sequencing reactions utilized fluorescent sequencing techniques with d-rhodamine and Big Dye terminator chemistry (Perkin Elmer-ABI, Foster City, CA) and were performed according to the standard protocols. The sequence identity between the two 3' fragments T1 (SEQ

ID NO:12) and T6 (SEQ ID NO:11) is 91.5%, suggesting the presence of two MFP1 genes in *Nicotiana tabacum*.

In a second round, the tobacco cDNA library was screened with a 1.0 kb 5' fragment of the LeMFP1 cDNA (p1-3; SEQ ID NO:10) (Meier et al., *Plant Cell* 8:2105-2115 (1996)). Two positive pfus were detected among approximately 600,000 pfus. Sequencing of the excised cDNAs (T2 (SEQ ID NO:13) and T3 (SEQ ID NO:14)) showed that they represented partial cDNAs, overlapping with T1 (SEQ ID NO:12) and T6 (SEQ ID NO:11) (Figure 3). Initial sequence analysis of the T2 (SEQ ID NO:13) and T6 (SEQ ID NO:11) cDNAs showed that they shared 445 bp of identical overlapping sequence. It was concluded the T2 (SEQ ID NO:13) and T6 (SEQ ID NO:11) cDNAs represent different portions of the same gene. The overlap of T3 (SEQ ID NO:14) and T1 (SEQ ID NO:12) is only 70 bp, and within this area, there is only a single base pair difference between T6 (SEQ ID NO:11) and T1 (SEQ ID NO:12).

In order to confirm that T3 (SEQ ID NO:14) and T1 (SEQ ID NO:12) were derived from the same gene, PCR primers (SEQ ID NO:25 and SEQ ID NO:26) were designed from the T3 (SEQ ID NO:14) and T1 (SEQ ID NO:12) sequences, that would allow the amplification of a 397 bp fragment, from a *Nicotiana tabacum* lambda ZAP cDNA library, overlapping both cDNAs. PCR reactions were carried out in a Perkin Elmer 9600 thermocycler. (Perkin Elmer, Foster City, CA). The thermocycler was programmed as follows: 2 min 96 °C denaturation cycle, was followed by 30 cycles of 94 °C, 45 sec; 55 °C, 45 sec; 72 °C, 90 sec, and ended with an 8 min 72 °C final extention cycle.

Cloning

Using restriction sites added to the primers, the PCR fragment was subsequently cloned into the XbaI/BamHI sites of pSK+ (Stratagene, La Jolla, CA). The sequence of the fragment PCR1 (SEQ ID NO:15) (Figure 3) was found to be 100% identical with both T1 (SEQ ID NO:12) and T3 (SEQ ID NO:14), confirming that these two cDNA fragments are derived from the same gene.

Figure 3 shows a schematic structure of the partial tobacco MFP1 cDNAs. T3, T1 and PCR1, shown as open boxes, represent overlapping fragments of the same gene (NtMFP1-1). T2 and T6, shown as filled boxes, represent overlapping fragments of a second gene (NtMFP1-2). The fragment used as a probe for the Southern blot (Figure 5) is indicated.

Confirmation Of The Presence Of Two Genes

The divergence between the two tobacco MFP1 cDNAs indicated that they were derived from two different genes. It has been previously shown that a single gene (SEQ ID NO:16) codes for MFP1 (SEQ ID NO:17) in tomato (LeMFP1)

(Meier et al., *Plant Cell* 8:2105-2115 (1996)). Applicants have additionally found that AtMFP1 is a single gene in *Arabidopsis* (data not shown). Based on these findings it was necessary to confirm whether MFP1 as also a single-copy gene in the two diploid progenitors of amphidiploid *Nicotiana tabacum*, *N. tomentosiformis* and *N. sylvestris*. In order to confirm this hypothesis the following procedure was applied.

For a Southern blot of *Nicotiana tabacum* genomic DNA, 20 μg aliquots of DNA were digested with various restriction enzymes, run out on 0.8% agarose gel, and were transferred to Immobilon N hydrophobic filters (Millipore, Bedford, MA). Hybridization conditions were essentially as described by Maniatis. The probe (SEQ ID NO:18) was prepared by purification of a 391 bp XhoI/SpeI fragment from the *Nicotiana tabacum* clone T3, as described above. The hybridization temperature was 65 °C. The probe (SEQ ID NO:18), shown in Figure 3, was labelled with ^{32}P by random prime method according to the manufacturers instructions (BRL, Gaitherburg, MD). Washes were performed at high stringency (0.1 x SSC, 0.1 % SDS at 65 °C).

In the region overlapping the probe, NtMFP1-1 (SEQ ID NO:1) contained a single XbaI site, whereas NtMFP1-2 (SEQ ID NO:3) contains no XbaI site. Neither of the two cDNAs contained an EcoRI site. Figure 5 shows the genomic organization of tobacco MFP1. Tobacco genomic DNA was digested with the indicated restriction enzymes, separated by agarose gel electrophoreses and hybridized in a genomic Southern blot with 391 bp Xho/Spe fragment from the Nicotiana tabacum cDNA clone T3 (shown in Figure 3). Abbeviations used are as follows: E, EcoRI; X, XbaI; E/X, EcoRI/XbaI; S, SspI; S/X, SspI/XbaI. The position of DNA size markers is indicated on the right. Two fragments were detected in the lane containing an EcoRI digest (approximately 3.7 kb and 2.7 kb) and three were seen in the lane containing an XbaI digest (approximately 8.0 kb, 7.5 kb and 5.0 kb). In the lane containing the EcoRI/XbaI double digest, the 3.7 kb EcoRI fragment appears to be cleaved by XbaI, leading to two smaller fragments of approximately 1.6 and 0.8 kb. This pattern is consistent with the presence of two genes, one of which contains an XbaI site in the region hybridizing to the probe. In addition, SspI and SspI/XbaI digests were analyzed. Again, one of the two bands detected in the SspI digest is cleaved in the SspI/XbaI double digest. The observed patterns are all consistent with the presence of two genes in the *Nicotiana tabacum* genome, represented by the two isolated cDNAs. These data indicate that, at least in tomato, tobacco and *Arabidopsis*, (data not shown), MFP1 is encoded by a single gene per diploid genome.

In summary, two distinct NtMFP1 cDNAS were isolated from tobacco and named NtMFP1-1 (SEQ ID NO:1) (containing T3 and T1) and NtMFP1-2 (SEQ ID NO:3) (containing T2 and T6). NtMFP1-1 (SEQ ID NO:1) is a full-length cDNA coding for a protein of 722 amino acids (SEQ ID NO:2). NtMFP1-1 (SEQ ID NO:1) and NtMFP1-2 (SEQ ID NO:3) have 77.0% and 78.9% identity to LeMFP1 (SEQ ID NO:16) on DNA level, respectively. The identity between the two tobacco sequences is 91.5%. NtMFP1-1 (SEQ ID NO:2) contains an open reading frame of 721 amino acids. It contains a short 69 bp 5' non-coding region preceding the ATG start codon. NtMFP1-2 (SEQ ID NO:4) contains an open reading frame of 398 amino acids and is not a full-length cDNA.

Table 1 lists the DNASTAR and BLAST comparison of Nt-MFP1-1 and Nt-MFP1-2 with a suit of public databases as well as the literature sequence for tomato (SEQ ID NO:16 and 17) and the MFP1 sequence isolated from Arabidopsis (http://genomewww.standard.edu/Arabidopsis/).

TABLE 1

Comparison of Nt-MPF1-1 and Nt-MPF1-2 Against Known Sequences

Gene/protein	% Identity* to Tomato		% Identity* to Arabidopsis		Similarity Identified	BLAST Algorithm	E value[1]	Citation
	AA[2]	NA[3]	AA[2]	NA[3]				
Nt-MFP1-1	77%	77%	42%	49%	(Y07861) MFP1 protein [*Lycopersicn esculentum*]	Xnr	0.0c	Meier et al., *Plant Cell* 8:2105-2115 (1996)
Nt-MFP1-2	79%	78%	42%	49%	(Y07861) MFP1 protein [*Lycopersicn esculentum*]	Xnr	e-147	Meier et al., *Plant Cell* 8:2105-2115 (1996)

*Comparison made using DNASTAR MEGALIGN, applying default parameters.

[1]Expect value. The Expect value estimates the statistical significance of the match, specifying the number of matches, with a given score, that are expected in a search of a database of this size absolutely by chance.

[2]AA is the abbreviation for amino acid sequence

[3]NA is the abbreviation for nucleotide sequence

EXAMPLE 5

Primary And Seconday Structure Analysis Of NtMFP1-1 And NtMFP1-2

Due to the small number of MFP1-like proteins discovered to date, it was advisable to confirm the identity of the present proteins through an analysis of secondary protein structure. Comparisons of the Nt-MFP1-1 and Nt-MFP1-2 proteins were made with the secondary structure of LeMFP1 isolated from tomato and AtMFP1 isolated from *Arabidopsis*.

The *Arabidopsis* genomic DNA sequence was accessed through the *Arabidopsis thaliana* Database (http://genomewww.standard.edu/Arabidopsis/). The deduced protein sequences of the MFP1 proteins were determined and compared using DNASTAR Lasergene software (DNASTAR, Inc., Madison, WI). Figure 4A shows the percent identical amino acids in pairwise comparisons of the four MFP1 proteins.

Based on the amino acid sequence identity NtMFP1-1 (SEQ ID NO:2) and NtMFP1-2 (SEQ ID NO:4) are most closely related. LeMFP1 (SEQ ID NO:17) is more closely related to the two tobacco MFP1s (NtMFP1-1 (SEQ ID NO:2) and NtMFP1-2 (SEQ ID NO:4)) (76% overall sequence identity) than to AtMFP1 (41% overall identity) reflecting the closer relationship of the two solanaeceous species.

Figure 4B shows the hydrophilicity and secondary structure analysis of NtMFP1-1 (SEQ ID NO:2), LeMFP1 (SEQ ID NO:17) and AtMFP1. The secondary structures of the proteins, hydrophilicity, α-helical, and coiled-coil regions were analyzed using DNASTAR PROTEAN software. AH indicates α-helical, CC indicates coiled-coil and HP indicates hydrophilicity plot. The hydrophobic domains are marked with open boxes. Like LeMFP1 (SEQ ID NO:17), NtMFP1 and AtMFP1 contain an extended α-helical, coiled-coil like domain and a shorter N-terminal, non-α-helical region that contains two hydrophobic domains. These structural features are extremely well conserved, despite a relatively low degree of identity on amino acid level in some areas. This is consistent with the more structural conservation of the positioning of polar and non-polar amino acids that is known from other filament-like coiled-coil proteins such as the nuclear lamins (McKeon et al., *Nature* 319:463-468 (1986)). The distance between the first and second hydrophobic domains is very similar in all three proteins (29 AA for tomato, 31 AA for tobacco, and 33 AA for *Arabidopsis* MFP1), indicating a functional relevance of the spacing between the two hydrophobic domains. The length of the N-terminal domain preceding the first hydrophobic domain varies between 56 AA for tomato, 61 AA for tobacco, and 72 AA for *Arabidopsis* MFP1. The common feature of this domain in all three

proteins is a relatively high content of serine and threonine residues (27% to 28%).

EXAMPLE 6

Composition Of cDNA Libraries And Identification Of cDNA Clones From Other Plant Species Encoding Homologs Of MFP1

cDNA libraries representing mRNAs from soybean or corn tissues were prepared. The characteristics of the libraries are described below in Table 2.

Table 2

cDNA Libraries from Plants	Library	Tissue
Soybean (*Glycine max*)	src3c.pk004.m1	8 day old root tissue inoculated with eggs of nematode
Corn (*Zea mays*)	p0118.chsab48r	stem tissue, night harvested
Rice (*Oryza sativa L.*, Nipponbare)	rca1n.pk022.a11	callus normalized

Soybean MFP1:

A soybean MFP1 cDNA was identified based on primary and secondary structure analysis. This sequence, from clone src3c.pk004.m1, came from a library prepared from 8 day old root tissue inoculated with eggs of cyst nematode for four days. This sequence contains 1164 base pairs of DNA (SEQ ID NO:19) encoding 388 amino acids (SEQ ID NO:20).

Comparison of this partial soybean MFP1 sequence (SEQ ID NO:20) with the sequences from tomato (LeMFP1; SEQ ID NO:17), tobacco (NtMFP1-1; SEQ ID NO:2), and *Arabidopsis* (AtMFP1) shows it to be 46.1, 45.9, and 40.7% identical to these sequences, respectively. In addition, secondary structure analysis of the partial soybean MFP1 (GmMFP1; SEQ ID NO:20) coded by SEQ ID NO:19 shows that it contains an extended α-helical, coiled-coil like domain as do the other MFP1 protein sequences. Results of a Southern blot experiment (not shown) suggest that the soybean MFP1 is encoded by a single copy gene.

Corn MFP1:

A corn EST sequence was identified as an MFP1 homolog from clone p0118.chsab48r. Secondary structure analysis of the corn MFP1 protein (SEQ ID NO:22) coded by SEQ ID NO:21 shows that it contains an extended α-helical, coiled-coil like domain as well as one of the hydrophobic domains at the N-terminus. Both features are indicative of MFP1 proteins (see Figure 4B).

Rice MFP1:

A rice EST, from clone rca1n.pk022.a11, was isolated which codes for an MFP1 protein. The identity of the rice EST was based on its high degree of identity to the corn MFP1 sequence (68%). This clone covers the C-terminal region that is most highly conserved between all MFP1 proteins identified. The rice MFP1 sequence (SEQ ID NO:23) codes for SEQ ID NO:24.

Table 3 lists the DNASTAR and BLAST comparison of the MFP1 sequences isloated from corn, soybean and rice with a suit of public databases as well as the literature sequence for tomato (SEQ ID NO:16 and 17) and the sequence isolated from Arabidopsis (http://genomewww.standard.edu/Arabidopsis/).

TABLE 3

Comparison of Corn, Rice and Soybean MPF1-1 Seqeunces Against Known Sequences

Gene/protein	% Identity* to Tomato		% Identity* to Arabidopsis		Similarity Identified	BLAST Algorithm	E value[1]	Citation
	AA[2]	NA[3]	AA[2]	NA[3]				
Corn MPF1	40%	40%	34%	48%	(Y07861) MFP1 protein [*Lycopersicn esculentum*]	Xnr	2e-94	Meier et al., *Plant Cell* 8:2105-2115 (1996)
Soybean MFP1	46%	51%	41%	48%	(Y07861) MFP1 protein [*Lycopersicn esculentum*]	Xnr	5e-76	Meier et al., *Plant Cell* 8:2105-2115 (1996)
Rice MFP1	39%	43%	31%	39%	(Y07861) MFP1 protein [*Lycopersicn esculentum*]	Xnr	6e-12	Meier et al., *Plant Cell* 8:2105-2115 (1996)

*Comparison made using DNASTAR MEGALIGN, applying default values

[1]Expect value. The Expect value estimates the statistical significance of the match, specifying the number of matches, with a given score, that are expected in a search of a database of this size absolutely by chance.

[2]AA is the abbreviation for amino acid sequence

[3]NA is the abbreviation for nucleotide sequence